Buổi 1: GIỚI THIỆU NGÔN NGỮ HỌC NGỮ LIỆU (Corpus Linguistics)

Nội dung

- Giới thiệu môn học
- Ngôn ngữ học ngữ liệu là gì?
- Tầm quan trọng của ngữ liệu
- Các loại ngữ liệu
- Giới thiệu một vài kho ngữ liệu

Giới thiệu môn học

- Xem đề cương

What do you learn?

- Khi hoàn thành môn học, sinh viên có thể:
  - Hiểu các khái niệm về ngôn ngữ học ngữ liệu
  - Hiểu các phương pháp đánh giá ngữ liệu
  - Biết một số bộ ngữ liệu phổ biến
  - Phân tích, thiết kế và xây dựng ngữ liệu cho một vài bài toán
  - Thử nghiệm, khảo sát và rút ra kết luận thông qua kết quả thử nghiệm
  - Thuyết trình đồ án, lắng nghe và đối thoại về kết quả thực hiện đồ án của những nhóm khác

Tài liệu tham khảo

- T. McEnery and A. Wilson. Corpus Linguistics. Edinburgh University Press, 2001.
- Michael Stubbs. Text and Corpus Analysis. Blackwell Publishers, 1996.

Trách nhiệm của sinh viên

- Học chăm chỉ
- Đọc tài liệu và tham gia thảo luận
- Nộp bài tập đúng hạn
- Tham dự đầy đủ các buổi học
- Có vấn đề không hiểu phải hỏi GV

What is a Corpus?

Ngữ liệu (corpus, số nhiều: corpora)

- Trong ngôn ngữ học và từ điển học, văn bản và các bài phát biểu (tiếng nói) có thể được xem là đại diện của một ngôn ngữ và thường được lưu trữ như một cơ sở dữ liệu điện tử.
  - Máy có thể đọc (computer-based)
  - Tin cậy (xảy ra một cách tự nhiên, không giả tạo)
  - Có thể làm mẫu (được thu thập từ nhiều nguồn)
  - Đại diện của một ngôn ngữ cụ thể hoặc nhiều ngôn ngữ
- Sinclair's (1996) definition:

  Một kho ngữ liệu là một tập hợp của các mẫu ngôn ngữ, được chọn lựa và sắp xếp theo các tiêu chí ngôn ngữ tường minh để được sử dụng như vật mẫu (sample) của ngôn ngữ.

What is a Corpus?

- Về cơ bản, bất kỳ tập nào có nhiều hơn 1 text thì được gọi là corpus

- Ngôn ngữ học ngữ liệu nghiên cứu về các ngữ liệu

(from The Oxford Companion to the English Language, ed. McArthur & McArthur, 1992)

Why Are Electronic Corpora Useful?

- Bộ sưu tập các ví dụ cho các nhà ngôn ngữ
- Nguồn dữ liệu cho các nhà từ điển học
- Tài liệu hướng dẫn dành cho giáo viên và người học ngôn ngữ
- Dữ liệu huấn luyện của các ứng dụng xử lý ngôn ngữ tự nhiên
  - Hệ thống nhận diện tiếng nói
  - Hệ thống gán nhãn từ loại, phân tích cú pháp
  - Hệ thống dịch máy thống kê, dịch dựa trên ví dụ
- "Big Data" là một corpus
  - Kỹ thuật phân tích "Big Data" tương tự các corpus khác

Ví dụ cho các nhà ngôn ngữ

- Cho ví dụ ngữ danh từ trong tiếng Việt
  - Một cô gái đẹp
  - Cô gái ném quả bóng là chị tôi

  -> "Cô gái ném quả bóng" có cấu trúc như một câu (S = NP VP), tuy nhiên trong ngữ cảnh này nó là một ngữ danh từ.

Why do Linguists need Corpora?

- Động từ *ăn* không thể theo sau bởi thức uống:
  - tôi *ăn* cơm
  - Tôi *ăn* nước cam -> sai
- Chomsky: động từ *perform* không thể có object là từ chỉ khối lượng
- Làm sao biết được cú pháp của từ *perform?*

This is why

- Tìm ”perform [nn1*]” trong corpus:
  - PERFORM MUSIC 4
  - PERFORM WORK 4
  - PERFORM SURGERY 3
  - PERFORM RESEARCH 2

many Continental musicians, and it can not be doubted that professional English singers often perform music which they have not had time to ” learn ” in any sense of

Not only do “ Saxtet ” perform music previously unassociated with the saxophone, but they include a selection of their own

→Có thể rút ra nguyên tắc sử dung của perform từ corpus?

Ví dụ cho các nhà từ điển học

- Từ ***line*** có bao nhiêu nghĩa? (30 nghĩa)

  Thống kê nghĩa của line dựa vào corpus

  – *telephone line, phone line,* ***line*** — (a telephone connection)

  – *the letter consisted of three short lines,* ***line*** — (text consisting of a row of words written across a page or computer screen)

  – *you must wait in a long line at the checkout counter,* ***line*** — (a formation of people or things one behind another)

  – …

Hướng dẫn học ngôn ngữ

- Trong tiếng Anh, nên dùng cái nào: *think about* hay *think on*?

- Nếu nghi ngờ, hỏi google:
  - 36,300,000 hits *think about*
  - 738,000 hits *think on*

Các kiểu kho dữ liệu

versus

Ngữ liệu đơn ngữ (mono-lingual)	Ngữ liệu đa ngữ (multi-lingual)
Ngữ liệu có mục đích đặc biệt (special-purpose), đặc thù miền (domain-specific)	Ngữ liệu với mục đích chung (general-purpose), quy mô lớn (large-scale)
Ngữ liệu tiếng nói (speech)	Ngữ liệu chữ viết (text)
Ngữ liệu không khái quát, được xây dựng phục vụ một yêu cầu cụ thể (ad-hoc)	Ngữ liệu cân bằng, có tính đại diện (representative)
Raw text	marked-up documents
Ngữ liệu chưa gán nhãn (unannotated)	Ngữ liệu đã được gán nhãn (annotated)
Web cũng được xem là corpus	

What does a corpus consist of?

- Một tập các file văn bản thông thường (ngữ liệu thô – raw corpus)
  - Ngữ liệu thô có thể được lưu dưới dạng xml/html/... (có thể bao gồm ngày public, chủ đề, thể loại (thơ, văn xuôi, ...), ...)
- Kho ngữ liệu có chú thích
  - Ngữ liệu có chú thích, có thể là nhãn từ loại, cấu trúc cú pháp, ...

The British National Corpus (BNC)

- 100 triệu từ British English (tiếng nói và chữ viết)
- Tiêu biểu cho British English cuối thế kỷ 20: cân bằng, có tính đại diện
- Gán nhãn từ loại (2 triệu từ được kiểm tra thủ công)

Written	Domain	Date	Medium
90%	Imaginative (22%) Arts (8%) Social science (15%) Natural science (4%) . . .	960-74 (2%) 1975-93 (89%) Unclassified (8%)	Book (59%) Periodical (31%) Misc. published (4%) Misc. un-pub (4%)
Spoken	Region	Interaction type	Context-governed
10%	South (46%) Midlands (23%) North (25%) . . .	Monologue (19%) Dialogue (75%) Unclassified (6%)	Informative (21%) Business (21%) Institutional (22%) . . .

Ngữ liệu tổng quát khác ngữ liệu chuyên dụng (General vs. specialized corpora)

- Ngữ liệu tổng quát: là một dự án khổng lồ, được xây dựng có tổ chức trong nhiều năm
- Ngữ liệu chuyên dụng (ví dụ: ngữ liệu bao gồm các cuộc hội thoại y khoa) có thể được xây dựng khá nhanh, dành cho các mục đích trước mắt, vì thế loại dữ liệu này phổ biến hơn
- Đặc điểm của ngữ liệu
  1. Máy có thể đọc, đáng tin cậy
  2. Làm mẫu, cân bằng và có tính đại diện

- Xu thế: đối với ngữ liệu chuyên dụng, đặc điểm (2) bị làm yếu, quan tâm sự hội tụ nhanh và kích thước lớn

  Hiện tựng ngôn ngữ hiếm chỉ xuất hiện trong tập dữ liệu lớn.

Giới thiệu ngữ liệu

- General corpus
  - Brown Corpus: text collection in the field of corpus linguistics, one million words, compiled from works published in the United States in 1961
- Parsed Corpora
  - Penn Treebank (WSJ, Brown, Chinese)
  - Czech Dependency Bank
  - Redwoods HPSG corpus of English

Ngữ liệu tiếng Việt